A TONE RECOGNITION FRAMEWORK FOR CONTINUOUS MANDARIN SPEECH

Lei He, Jie Hao

Toshiba (China) Research and Development Center
W2, Oriental Plaza, Beijing, 100738, P.R.China
{helei, haojie}@rdc.toshiba.com.cn

10.21437/Interspeech.2006-441

ABSTRACT

In this paper, we present a tone recognition framework for continuous Mandarin speech. To model the variations of F0 pattern caused by co-articulation and phonetic effects, a set of discriminating features are extracted: 1) outlined features from the F0 contours of target syllable and neighboring syllables are combined; 2) contextual tone information is utilized within an iterative process; 3) phonetic information from target and neighboring syllables is incorporated. These features are put into a decision tree for tone classification, which follows an HMM-based toneless decoder. The results in 5-tone recognition experiments show more than 40% relative error rate reduction against the baseline local outlined features. Moreover, the proposed method obviously outperforms HMM-based tone model in speaker-independent evaluation.

Index Terms: speech recognition, tone recognition, feature selection.

1. INTRODUCTION

It is well known that Mandarin is a tonal language, in which each character is pronounced as a single syllable and is associated with a lexical tone. There are four basic tones and a neutral tone in Mandarin. The basic tones can be represented by four distinctive F0 contour patterns, whereas the neutral tone is highly dependent on the preceding tone, without a specific F0 pattern [1].

In the viewpoint of speech recognition, lexicon tones are important to distinguish a number of homophonic words in Mandarin. Moreover, tone information can improve the recognition accuracy from a digital string recognition task [2] to an LVCSR system [3]. Therefore, many researches on automatic tone recognition have been pursued in the past decades. Most of these methods can be classified into two approaches. One is to extend the toneless acoustic model to tonal model (such as tonal phonemes, tonal Finals, etc.) [4][5]. In general, cepstral features and tonal features (e.g., F0 and corresponding derivatives) from time frames are used to train HMM to form tonal models. In this approach, tone recognition is performed simultaneously (one-pass approach). The disadvantage of this approach is the increased model size and decoding time. The other approach is to model the tone pattern separately: 1) tonal features of time frames can be used to train an HMM for each tone [6]; 2) outlined features can be extracted from F0 contours to model the tone pattern [4][6], and then discriminated by static classifiers, such as ANN [7], GMM [8], decision tree [9], etc. Here, a toneless recognizer can produce the syllable boundaries for outlined features extracting, which results in a two-pass framework.

Previous studies have shown satisfying tone recognition performance for isolated syllables [3][10]. Nevertheless, the tone pattern in continuous speech usually deviates from the typical one due to complex phonetic and linguistic effects. Thus tone recognition in continuous speech is still a challenge. Some relevant studies have been made in recent years. The fractionized tone models were trained to describe tone pattern in continuous speech [4]. In [6], the underlying segment of F0 contour was detected and then used to model the tone pattern. The contextual features from neighboring syllables were extracted [7][8].

In this paper, we carry out our research in a two-pass framework, aiming at finding a solution to tone recognition for continuous Mandarin speech under the limitations of the storage of acoustic model and computing resources. The basic idea is to select discriminating features to model the tone pattern variation in continuous speech. First of all, the subsection outlined features from F0 contour of target syllable are extracted to capture the main structure of tone pattern. Next, the contextual outlined features and the neighboring tone information are combined to describe the co-articulation effects. Moreover, local and neighboring phonetic information is explicitly incorporated. Here, a toneless HMM-based decoder is used to produce the phonetic information and syllable boundaries. Whereas the neighboring tone information is utilized based on an iterative processing. At last, we use a decision tree to fuse all these features.

This paper is organized as follows. Section 2 describes the system framework and the subsection outlined features. The contextual and phonetic feature selection is presented in Section 3. The experiments are proposed in Section 4, followed by conclusions in last Section.

2. SYSTEM

2.1 System framework

Figure 1 shows the system framework of proposed method. Firstly, toneless recognition is performed using a toneless HMM, with the outputs of toneless candidates and relevant segmentation information. In the second step of tone recognition, all tonal features are extracted from three sources: 1) F0 contours of utterances, 2) toneless candidates and segmentation information given by toneless

recognizer, 3) contextual tone assumption from previous iteration of tone recognition. We follow the normalized short time autocorrelation algorithm presented in [11] for F0 Extraction. Especially, the correlation coefficient of each frame is used to represent the voicing level (VL) explicitly, and is adopted as an additional feature. All these features are finally combined and fed into a decision tree. Basically, all features can be simply calculated according to the critical resource limitation.

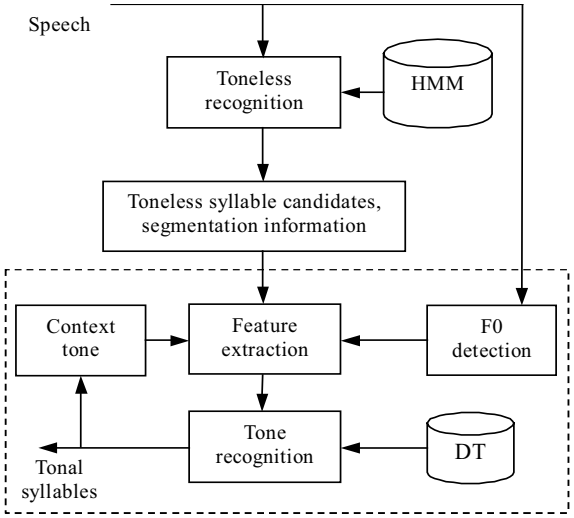

Figure 1. System block diagram

2.2 Subsection outlined features

It has been shown that the outlined features of F0 contours are efficient in tone recognition [4]. Here, the tone pattern is captured by several outlined features that are not so sensitive to F0 extraction errors as the details of F0 contours.

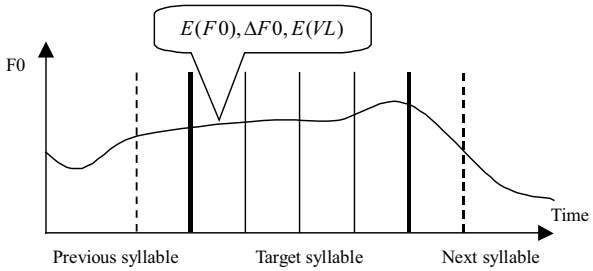

Figure 2. Subsection outlined features definition

In this paper, the voiced phase of a syllable is divided into four segments as shown in figure 2. In each segment, the average F0 value, average voicing level and movement of F0 value are extracted. Plus the duration of syllable, 13-dimension features are used to model the tone pattern of a syllable as the baseline in following experiments. Here, the parameter of voicing level gives accuracy improvement in tone recognition according to the comparison experiments. Additionally, the linear segmentation gives better performance than some other complicated methods in experiments, which is similar to the conclusion in [4].

3. FEATURE SELECTION

We have shown that tone recognition in continuous speech is not a trivial task due to the complex F0 pattern variation. The most important issue in tone recognition is to find appropriate features to efficiently model such variation. In this section, we focus on two main effect factors of tone pattern change: co-articulation effects and phonetic effects, and investigate three kinds of features to improve the tone discrimination in continuous speech.

3.1 CONTEXTUAL FEATURES EXPANSION

In general, due to the effects of co-articulation, F0 contours of a syllable are quite different in different context in continuous speech, which results in the difficulty of reliable modeling. In an extreme case, different tones have similar F0 contours, leads to serious overlaps in feature space. Therefore, a natural idea is to extent the features to the neighboring syllables besides the local features of processing one, to mitigate such overlaps. In this paper, the features from last segment of previous syllable and first segment of next syllable are combined, in expectation of reinforcing the tone discrimination (Illustrated in figure 2).

3.2 CONTEXTUAL TONE INFORMATION

Previous section introduces the contextual outlined features to describe co-articulation effects on tone pattern variation. Here, more explicit features are adopted. Among all co-articulation effects derived from continuation of articulation process, the neighboring tone types are most remarkable for tone pattern variation, which is also known as sandhi rules [2]. In some cases, the tone pattern totally changes to another one. For example, the third tone followed by another third tone usually shows similar pattern with a general second tone. Therefore, *a priori* contextual tone information should be helpful to tone recognition in continuous speech. Since such information is not available in recognition, an iterative process is introduced, in which the output tones in previous iteration are fed back as the input of current iteration (Illustrated in figure 1).

3.3 PHONETIC CATEGORY INFORMATION

Besides co-articulation effects, the phonetic effects on tone pattern variation are also remarkable. In fact, the F0 contour of a syllable is decided not only by the canonical tone type, but also by the local and contextual phonetic content. An example is that the voiced or unvoiced Initials will directly affect the tone pattern in a whole syllable. Additionally, the neutral tone occurs more frequently in a subset of syllables. Therefore, the phonetic information of current and neighboring syllables should also improve the tone discrimination. In this paper, all phonetic units (Initials and Finals) are clustered into 7 classes according to corresponding phonetic attributes. The class ID of each unit is used to represent phonetic information explicitly. There are two reasons to use unit classes instead of certain

unit IDs. Firstly, the units in one class usually show similar effects on tone pattern. Secondly, since such phonetic information is produced by toneless recognition process, the interference of recognition errors is inevitable. Obviously, the use of unit classes can greatly reduce such negative effects.

4. EXPERIMENTS

4.1. EXPERIMENTAL ENVIRONMENT

In experiments, two speech databases are used. One is a speaker-dependent database that consists of 3,160 sentences uttered by a female speaker. The F0 values, syllable boundaries and tone identities of this corpus are manually checked. We use it for tonal features selection to reduce the effects of segmentation and pitch extraction errors. The other is a speaker-independent corpus, which is selected from the Mandarin speech database in National High Technology Development Project (863project) of China. This corpus is used to farther investigate the system performance in entire two-pass recognition framework. The canonical tone identities in transcription are directly adopted, without manually checking. Here, speeches from 76 speakers are used for training with other 10 speakers for testing. Some details of database arrangement are in table 1.

Table 1. Databases arrangement

	Speaker dependent		Speaker independent	
	Sentences	Syllables	Sentences	Syllables
Train	2,707	51,943	39,504	509,598
Test	453	7,364	5,206	67,165

Additionally, the first step toneless HMM-based recognizer is a standard Viterbi decoder, which performs a free loop toneless syllable recognition task without LM. The features include 12-dimension MFCC and 1-dimension log energy, with corresponding first- and second-order derivatives, totally 39-dimension vector. The model units include 101 right context dependent Initials and 38 context independent Finals, each is described by 3-state left-to-right HMM with 16 Gaussians per state. In the second-pass tone recognition, we choose a decision tree C4.5, which is suitable for both the numerical and categorical features, with acceptable computing cost.

4.2. EXPERIMENTS ON FEATURE SELECTION

Based on the speaker dependent database, detailed feature selection experiments are carried out. Here, we only focus on the second pass tone recognition, and use the manually segmentation and correct phonetic information.

Firstly, the voicing level parameters from segments of F0 contours show efficiency in tone recognition with 8.43% error rate reduction (ERR). Here, only the outlined features of processing syllable are used, plus the duration of syllable. This feature set including VL parameter is adopted as the baseline in following experiments.

Secondly, comparison experiments are carried out after combining the outlined features from neighboring syllables. According to figure 3 (TN means the N-th tone, T5 is the neutral tone), contextual outlined features from neighboring syllable give obvious accuracy improvement. Using both left and right expansion (LR-ex in figures), 21.77% ERR is achieved. Additionally, we can also notice that the left contextual features (L-ex) contribute more discrimination than the right ones (R-ex), which demonstrates that the co-articulation of speech production is more carryover than anticipation [8].

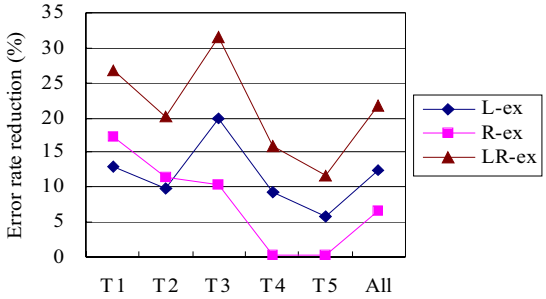

Figure 3. Experiment on contextual outlined features

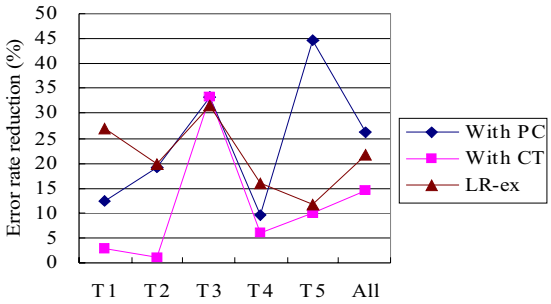

Figure 4. Experiment on three reinforced features

Next, the contextual tones (CT) and phonetic category (PC) information are added into the baseline system respectively. Here, CT includes the tones of preceding and following syllables given by previous iteration. In experiments, only two iterations are used to reduce the computing cost. As to the phonetic information, the class IDs of previous Final, current and following Initials and Finals, totally 5-dimension features are added. Plus the outlined feature expansion (left and right), the effects of these three kinds of features can be seen in figure 4 respectively. All three additional feature sets show obvious accuracy improvement in tone recognition. Especially, the performance improvement given by phonetic information directly demonstrates its discrimination in tone classification, which obviously outperforms other two features for neutral tone.

4.3. EXPERIMENTS ON FEATURE FUSION

Previous section respectively shows the efficiency of proposed three kinds of features for tone discrimination. In

this experiment, all above features are combined and evaluated. The results in table 2 show great accuracy improvement both in speaker-dependent and speaker-independent tasks. Here, the speaker-independent evaluation is based on the entire two-pass recognition scheme. The top 1 toneless candidate and corresponding segmentation information are produced by the HMM-based toneless recognizer, and then used for tonal features extraction.

Table 2. Experiments with combining features

	Tone recognition accuracy	
	Speaker-dependent	Speaker-independent
1. Baseline	80.96%	71.25%
2. Proposed method	89.06%	83.13%
ERR from 1. to 2.	42.54%	41.32%

Next, the HMM-based tone models are compared in speaker-independent task. The feature vector includes 39-dimension cepstral mentioned in section 4.1 and 4-dimension pitch features: F0 of each frame and corresponding first- and second-order derivatives, plus voicing level of each frame. The model topology is same as that in toneless recognizer. Two kind of tone models are tried here: 1) phone independent tone model. 5 models are used and each for one tone. 2) Phone dependent tone model. 101 toneless Initials and 168 tonal Finals are trained, nearly twice the toneless HMM. According to the results in table 3, the phone independent HMM-based tone model is not an appropriate solution to tone recognition in continuous speech. Phone dependent tone model greatly improves the performance. However, the proposed method still obvious outperforms the phone dependent tone model in experiments.

Table 3. Comparison with HMM-based tone model in speaker independent task

	Model	Tone accuracy
Proposed method	DT	83.13%
Phone dependent tone models (269 models)	HMM	76.48%
Phone independent tone models (5 models)	HMM	53.53%

Table 4. Tone accuracy (%) in speaker independent task

	T1	T2	T3	T4	T5
Baseline	78.4	85.75	43.9	80.9	16.5
Proposed method	86.4	88.0	71.0	88.7	53.8
Phone dependent tone models	82.2	84.3	70.6	79.7	58.3

At last, from detailed results in table 4, we notice that local outlined features give very poor accuracy for neutral tone. After reinforcing contextual and phonetic features, performance improves remarkably, but is still lower that the phone dependent tone model. Such phenomenon is coherent to the fact that neutral tone is hardly to be modeled by outlined features of F0 contours due to the non-stable tone pattern.

5. CONCLUSIONS

In this paper, sets of features are investigated for tone recognition, based on the analysis of tone pattern change in continuous Mandarin speech. For co-articulation effects, the contextual features from neighboring syllables reinforce the local outlined features. Meanwhile, the neighboring tone information is utilized within an iterative process. For phonetic effects, the phonetic category information from target syllable and neighboring syllables is incorporated. All features are fused with a decision tree classifier following an HMM based toneless recognizer, which forms a two-pass tone recognition framework. Experiments show great accuracy improvement in 5-tone recognition task with combination of these features against general local outlined features. Moreover, proposed method also outperforms HMM-based phone dependent tone model in experiments.

6. REFERENCES

[1] Y. Chao, “A Grammar of Spoken Chinese”, University of California Press, Berkeley.

[2] C. Wang, S. Seneff, “A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition”, *Proc. ICSLP*, pp.535-538, 1998.

[3] H. Hon, B, Yuan, Y. Chow, et al, “Toward Large vocabulary Mandarin Chinese speech recognition”, *Proc. ICASSP*, Vol. I, pp.545-548, 1994.

[4] Y. Tian, J. Zhou, M. Chu and E. Chang, “Tone recognition with Fractionized models and outlined features”, *Proc. ICASSP*, Vol. I, pp.105-108, 2004.

[5] C. Huang, Y. Shi, J. Zhou, et al, “Segmental tonal modeling for phone set design in Mandarin LVCSR”, *Proc. ICASSP*, Vol. I, pp.901-904, 2004.

[6] J. Zhang, K. Hirose, “Tone nucleus modeling for Chinese lexical tone recognition”, *Speech Communication*, Vol. 42, pp.447-466, 2004.

[7] S. Chen, Y. Wang, “Tone recognition of continuous Mandarin speech based on Neural Network”, *IEEE. Trans. Speech and Audio Processing*, Vol. 3, No. 2, pp.146-150. March 1995.

[8] J. Zhang, S. Nakamura, K. Hirose, “Efficient tone classification of speaker independent continuous Chinese speech using anchoring based discriminating features”, *Proc. ICSLP*, 2004.

[9] P. Wong, M. Siu, “Decision tree based tone modeling for Chinese speech recognition”, *Proc. ICASSP*, Vol. I, pp.905-908, 2004.

[10] W. Yang, J. Lee, Y. Chang, et al, “Hidden Markov model for Mandarin lexical tone recognition”, *IEEE. Trans. ASSP,* Vol. 36, No. 7, pp.988-992, July 1988.

[11] K. Hirose, H. Fujisaki, S. Seto, “A scheme for pitch extraction for speech using autocorrelation function with frame length proportional to the time lag”, Proc. ICASSP, Vol. I, pp. 149-152, 1992.